Interdisciplinary Sciences: Computational Life Sciences (2019) 11:559–574
https://doi.org/10.1007/s12539-019-00343-w

ORIGINAL RESEARCH ARTICLE

Estimation of Probability Distribution and Its Application in Bayesian Classification and Maximum Likelihood Regression

Hao Dai[1,3] · Wei Wang[2] · Qin Xu[1] · Yi Xiong[1] · Dong-Qing Wei[1,4]

Received: 21 November 2018 / Revised: 18 May 2019 / Accepted: 5 July 2019 / Published online: 17 July 2019

Abstract

Nonparametric estimation of cumulative distribution function and probability density function of continuous random variables is a basic and central problem in probability theory and statistics. Although many methods such as kernel density estimation have been presented, it is still quite a challenging problem to be addressed to researchers. In this paper, we proposed a new method of spline regression, in which the spline function could consist of totally different types of functions for each segment with the result of Monte Carlo simulation. Based on the new spline regression, a new method to estimate the distribution and density function was provided, which showed significant advantages over the existing methods in the numerical experiments. Finally, the density function estimation of high dimensional random variables was discussed. It has shown the potential to apply the method in classification and regression models.

Keywords Distribution function estimation · Density function estimation · Spline regression · Smoothing spline · Bayesian classification · Maximum likelihood regression

1 Introduction

Estimation of cumulative distribution function (CDF) and probability density function (PDF) to random variables is a classical and basic problem in statistics, which is essential to describe some random phenomena and has significant application in signal processing [1], pattern recognition [2], machine learning [3] and so on. With the known distribution of the continuous random variable, such as Gaussian, Rayleigh, log-normal or exponential distribution, CDF and PDF can be estimated with the maximum likelihood estimation and Bayes estimation [4]. But nonparametric approach will be employed here if the distribution is not well assumed.

To estimate PDF more exactly in nonparametric approach is still a challenging problem to researchers. As the most widely used method, kernel density estimation is proposed by Rosenblatt [5] in 1956 and Parzen [6] in 1962. Many discussions have been performed to further implement such method via optimize the kernel function and bandwidth, e.g., based on the normal distribution, normal scale rules is proposed by Silverman [7] to determine the best bandwidth; Over smoothed bandwidth selection rules from Terrell [8] is more flexibility and larger application; Alexandre [9] provided iterative algorithm used when solve the equation and Plug-In estimator to give the best bandwidth corresponding to the least mean integrated square error. For the large samples with high complexity, fast Parzen density estimation by Jeon and Landgrebe [10], weighted Parzen window by Babich and Camps [11], optimally condensed data samples by Girolami and He [12], etc. are based on the subset of the large sample to reduce the running time without reduce the accuracy. What's more, some other approaches have also been proposed to estimate PDF. Such as the sum of gamma densities [13] or a sum of exponential random variables [14] was used to substitute the kernel function to express the PDF in different fields, and orthogonal series [15], Haar's series

✉ Dong-Qing Wei
dqwei@sjtu.edu.cn

[1] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

[2] School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

[3] Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China

[4] Shanghai Jiao Tong University, Room 4-321, Life Science Building, 800 Dongchuan Road, Minhang District, Shanghai 200240, China

[16], wavelets [17]. Recently, some methods based on characteristic function [18, 19] were presented. However, all these methods are based on the series to express the PDF, in which the complexity of function would be increased with the sample size. And the accuracy of estimation was determined by the form of series, which meant one method was just suitable for some certain distributions. However, prior the process of estimation, there are little information available for us to the distribution, it will be very hard to have the proper series introduced in the estimation.

Spine functions have been widely applied in interpolation [20], smoothing of observed data [21], regression [22] and PDF estimation [23–25]. Inspired with the characteristics of spline function, which is a continuous function piecewise-defined by polynomial functions and possesses a high degree of smoothness where the pieces connect, and to overcome such shortness, a new method to estimate CDF or PDF is introduced here in this paper. Spline Not as previous methods, in our proposed method of spline regression, the spline function was not always defined by polynomial functions or B-splines, but could be set freely and consisted of totally different types of functions in each segment. With the method here, a new method to estimate CDF and PDF was introduced, which showed advantages in these aspects:

1. The PDF is expressed by piecewise functions instead of series. The estimated accuracy increases with the sample size, but the complexity of function does not increase.
2. The method is suitable for most types of continuous distributions, and the form of spline function and other parameters does not need to be updated unless the distribution is quite special.
3. The estimation is accurate for most types of distributions and is superior to kernel density estimation.
4. The PDF is always smooth and is not influenced by parameters.
5. The values of estimated CDF are less than 1, positive and monotone increasing. The values of estimated PDF are positive and the integration of PDF is about 1.
6. It is easy to find a subset from the large sample to reduce the running time and get similar accuracy simultaneously.

The paper is organized as below in the following sections, the new spline regression is introduced first, and then the application of proposed approach is described in the estimation of CDF and PDF. After that, comprehensive numerical experiments with Monte-Carlo simulation were made to illustrate the characteristic and advantage. At last, the PDF estimation of high dimensional random variables is discussed, and its potential application in classification and regression models is presented.

2 Method

Let $F(x)$ and $f(x)$ denote the CDF and PDF of random variable X, respectively, $y(x)$ and $y^{'(x)}$ denote the estimated CDF and PDF, respectively.

With ascending sorted samples from random variable X,

$x_1, x_2, \ldots, x_n$, where $x_1 \le x_2 \le \cdots \le x_n$

the CDF of x_i is $F(x_i) = P(X \le x_i) \approx i/(n+1)$, which means the probability of event $\{X \le x_i\}$ is almost to be $i/(n+1)$. If we let $y_i = i/(n+1)$, the data points $(x_i, y_i) i = 1 \cdots n$ can be fitted with spline regression. Then the PDF of x_i can be estimated with the one order deviation of $y(x)$, noted as $y^{'(x)}$. Instead of spline interpolation, spline regression is used in the paper, which means that $F(x_i)$ is not always equal to y_i.

To avoid the large error, which is resulted from the process of derivation, transformation of random variables is employed here in the paper.

2.1 Spline Regression

Inspired with the characteristics of spline function, A new method of spline regression is introduced here, in which the spline function can be set freely and the basis functions may be totally different for each segment.

The spline function is defined as

$$y(x) = \sum_{j=1}^{v} a_{ij}\varphi_{ij}(x) \quad \text{with} \quad x \in [s_i, s_{i+1}], \quad i = 1, 2, \ldots, u. \tag{1}$$

There are u segments in this function. For each segment to the interval $x \in [s_i, s_{i+1}], i = 1, 2, \ldots, u$, v basis functions $\varphi_{i1}(x), \ldots, \varphi_{iv}(x)$ are set here, which are smooth for each segment and with nonzero derivate for each knot $s_2, s_3, \ldots, s_u$ for their any order derivative. With the request of smoothness to the spline function, the following constrained conditions are introduced here:

$$\begin{cases} \sum_{j=1}^{v} a_{ij}\varphi_{ij}(s_{i+1}) = \sum_{j=1}^{v} a_{i+1,j}\varphi_{i+1,j}(s_{i+1}) & i = 1, \ldots, u-1 \\ \sum_{j=1}^{v} a_{ij}\varphi_{ij}^{(k)}(s_{i+1}) = \sum_{j=1}^{v} a_{i+1,j}\varphi_{i+1,j}^{(k)}(s_{i+1}) & i = 1, \ldots, u-1k = 1, \ldots, v-2, \end{cases} \tag{2}$$

where $\varphi_{ij}^{(k)}(x)$ is the kth order derivative of $\varphi_{ij}(x)$.

For these $u \cdot v$ parameters $a_{11}, \ldots, a_{uv}$ in the spline function, $(u-1) \cdot (v-1)$ linear constrained equations should be met, which means that there are $u + v - 1$ free variables in total, noted as $I_1, I_2, \ldots, I_{u+v-1}$. As equation set here is homogeneous linear equations, based on the form of solutions, all a_{ij} can be rewrote as

$$a_{ij} = \sum_{k=1}^{u+v-1} b_{ijk} I_k. \tag{3}$$

The spline function will be available if we can get the values of all b_{ijk} and I_k, which will be basis to the value of a_{ij}.

(1) Values of b_{ijk}

After transposition of the constrained equation set, we can get

$$\begin{pmatrix} \varphi_{i+1,2}(s_{i+1}) & \varphi_{i+1,3}(s_{i+1}) & \cdots & \varphi_{i+1,v}(s_{i+1}) \\ \varphi'_{i+1,2}(s_{i+1}) & \varphi'_{i+1,3}(s_{i+1}) & \cdots & \varphi'_{i+1,v}(s_{i+1}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{(v-2)}_{i+1,2}(s_{i+1}) & \varphi^{(v-2)}_{i+1,3}(s_{i+1}) & \cdots & \varphi^{(v-2)}_{i+1,v}(s_{i+1}) \end{pmatrix} \begin{pmatrix} a_{i+1,2} \\ a_{i+1,3} \\ \vdots \\ a_{i+1,v} \end{pmatrix} = \begin{pmatrix} -\varphi_{i+1,1}(s_{i+1}) & \varphi_{i1}(s_{i+1}) & \cdots & \varphi_{iv}(s_{i+1}) \\ -\varphi'_{i+1,1}(s_{i+1}) & \varphi'_{i1}(s_{i+1}) & \cdots & \varphi'_{iv}(s_{i+1}) \\ \vdots & \vdots & \ddots & \vdots \\ -\varphi^{(v-2)}_{i+1,1}(s_{i+1}) & \varphi^{(v-2)}_{i1}(s_{i+1}) & \cdots & \varphi^{(v-2)}_{iv}(s_{i+1}) \end{pmatrix} \begin{pmatrix} a_{i+1,1} \\ a_{i1} \\ \vdots \\ a_{iv} \end{pmatrix}, \quad (4)$$

in which $i = 1, 2, \dots, u-1$.

Take note that if we know the values of $a_{11}, a_{12}, \dots, a_{1v}, a_{21}, a_{31}, \dots, a_{u1}$, all a_{ij} will be derived accordingly with Eq. (4), all these $u+v-1$ variables $a_{11}, a_{12}, \dots, a_{1v}, a_{21}, a_{31}, \dots, a_{u1}$ will be set as free variables.

If some $I_k = 1$ and all others are 0 for the equation $a_{ij} = \sum_{k=1}^{u+v-1} b_{ijk} I_k$, $a_{ij} = b_{ijk}$. b_{ijk} is derived as below: let one of $a_{11}, a_{12}, \dots, a_{1v}, a_{21}, a_{31}, \dots, a_{u1}$ be 1 and the others are 0, substitute them into Eq. (4), we can get all other a_{ij} and all b_{ijk}.

(2) Values of I_k

For each $x_i \in [s_k, s_{k+1}]$,

$$y(x_i) = \begin{pmatrix} a_{k1} & a_{k2} & \cdots & a_{kv} \end{pmatrix} \begin{pmatrix} \varphi_{k1}(x_i) \\ \varphi_{k2}(x_i) \\ \vdots \\ \varphi_{kv}(x_i) \end{pmatrix} = \begin{pmatrix} I_1 & I_2 & \cdots & I_{u+v-1} \end{pmatrix} \begin{pmatrix} b_{k11} & b_{k21} & \cdots & b_{kv1} \\ b_{k12} & b_{k22} & \cdots & b_{kv2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1,u+v-1} & b_{k2,u+v-1} & \cdots & b_{kv,u+v-1} \end{pmatrix} \begin{pmatrix} \varphi_{k1}(x_i) \\ \varphi_{k2}(x_i) \\ \vdots \\ \varphi_{kv}(x_i) \end{pmatrix}. \quad (5)$$

Let

$$B_{im} = \sum_{j=1}^{v} b_{kjm} \varphi_{kj}(x_i).$$

Substitute it into the above equation, and then $y(x_i)$ can be represented as

$$y(x_i) = \sum_{m=1}^{u+v-1} B_{im} I_m. \quad (6)$$

According to the constrained equations, the $v-2$ order derivative of spline function is continuous, but the $v-1$ order derivative is not continuous. Nevertheless, it is hoped there is stronger smoothness for the spline function, with De Boor's smoothing spline [26] and define as

$$G = \frac{1}{n} \sum_{i=1}^{n} \left[y^{(v-1)}(x_i) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{v} a_{ij} \varphi_{ij}^{(v-1)}(x_i) \right]^2 .$$

Let

$$A_{im} = \sum_{j=1}^{v} b_{kjm} \varphi_{kj}^{(v-1)}(x_i).$$

Then

$$G = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{m=1}^{u+v-1} A_{im} I_m \right)^2 . \quad (7)$$

We should minimize both the value of G and the sum of squared residues, so we define

$$\begin{aligned} Q &= \sum_{i=1}^{n} \left[y_i - y(x_i) \right]^2 + \sigma G \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{u+v-1} B_{im} I_m \right)^2 + \frac{\sigma}{n} \sum_{i=1}^{n} \left(\sum_{m=1}^{u+v-1} A_{im} I_m \right)^2 , \end{aligned} \quad (8)$$

where σ is a parameter called smooth factor that we should set.

Based on the least square method,

$$\begin{aligned} \frac{\partial Q}{\partial I_t} = & -2 \sum_{i=1}^{n} B_{it} \left(y_i - \sum_{m=1}^{u+v-1} B_{im} I_m \right) \\ & + \frac{2\sigma}{n} \sum_{i=1}^{n} A_{it} \left(\sum_{m=1}^{u+v-1} A_{im} I_m \right) = 0 \quad t = 1, \dots, u+v-1. \end{aligned}$$

After transposition

$$\begin{aligned} \sum_{i=1}^{n} B_{it} y_i = & \sum_{i=1}^{n} B_{it} \left(\sum_{m=1}^{u+v-1} B_{im} I_m \right) \\ & + \frac{\sigma}{n} \sum_{i=1}^{n} A_{it} \left(\sum_{m=1}^{u+v-1} A_{im} I_m \right) \quad t = 1, \dots, u+v-1. \end{aligned}$$

Update that into matrix form, it can be rewrote to

$$\left(B^{\mathrm{T}} B + \frac{\sigma}{n} A^{\mathrm{T}} A \right) I = B^{\mathrm{T}} Y, \quad (9)$$

where $B = (B_{im})_{n \times (u+v-1)}$, $A = (A_{im})_{n \times (u+v-1)}$, n is the sample size, σ is smooth factor, $I = \begin{pmatrix} I_1 & I_2 & \cdots & I_{u+v-1} \end{pmatrix}^{\mathrm{T}}$ and $Y = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^{\mathrm{T}}$.

The value of I_k can be derived with the following steps:

1. For each x_i, calculate the values of $B_{im} = \sum_{j=1}^{v} b_{kjm}\varphi_{kj}\left(x_i\right)$ and $A_{im} = \sum_{j=1}^{v} b_{kjm}\varphi_{kj}^{(v-1)}\left(x_i\right)$, $i = 1, \dots, n \quad m = 1, \dots, u + v - 1$.
2. Set appropriate smooth factor σ and solute the equation $\left(B^{\mathrm{T}}B + \sigma/n \cdot A^{\mathrm{T}}A\right)I = B^{\mathrm{T}}Y$, values of $I_1, I_2, \dots, I_{u+v-1}$ will be derived accordingly.
3. All parameters of the spline function can be derived with $a_{ij} = \sum_{k=1}^{u+v-1} b_{ijk}I_k$.

Take note that we suppose all matrices above are full rank. If some matrix is not full rank, another group of basis functions or knots will be used.

2.2 Transformation of Random Variable *X*

In the case of that some CDFs are not easily estimated with general spline function, such as, $F(x) = \sqrt{x}, x \in [0, 1]$, in which $\lim\limits_{x \to 0} f(x) = +\infty$. CDF cannot be estimated with polynomial spline function to fit the data. However, if we set $\hat{x} = \ln x$, then $F(x) = e^{\hat{x}/2}, \hat{x} \in (0, +\infty)$, which is much easier to be estimated.

For random variable X, set $\hat{X} = \psi(X)$ in which ψ is a monotone increasing and analytic function, and let $\hat{F}(\hat{x})$ and $\hat{f}(\hat{x})$ are the distribution function and density function of $\hat{X}$, respectively.

Then

$$\begin{aligned} F(x) &= P(X \le x) = P\left(\psi^{-1}\left(\hat{X}\right) \le x\right) \\ &= P\left(\hat{X} \le \psi(x)\right) = P\left(\hat{X} \le \hat{x}\right) = \hat{F}(\hat{x}) \end{aligned}$$

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \frac{\mathrm{d}\hat{F}(\hat{x})}{\mathrm{d}\hat{x}} \cdot \frac{\mathrm{d}\hat{x}}{\mathrm{d}x} = \hat{f}(\hat{x})\psi'(x). \tag{10}$$

Using spline function to fit the data points $\left(\hat{x}_i, y_i\right) i = 1 \dots n$ in which $\hat{x}_i = \psi\left(x_i\right)$, we can get $\hat{F}(\hat{x})$ and $\hat{f}(\hat{x})$, and then we can get $F(x)$ and $f(x)$ based on the above equations.

In this paper, we transformed the random variables based on the following steps:

For ordered samples: $x_1, x_2, \dots, x_n\left(x_1 \le x_2 \le \cdots \le x_n\right)$, noted the the a quantile as x_{an}.

Step 1:

If $\frac{x_{0.02n}-x_1}{x_{0.2n}-x_1} < 0.02$ and $\frac{x_n-x_{0.98n}}{x_n-x_{0.8n}} \ge 0.02$, $\psi_1 = \ln\left(x - x_1\right)$.

If $\frac{x_{0.02n}-x_1}{x_{0.2n}-x_1} \ge 0.02$ and $\frac{x_n-x_{0.98n}}{x_n-x_{0.8n}} < 0.02$, $\psi_1 = -\ln\left(x_n - x\right)$.

If $\frac{x_{0.02n}-x_1}{x_{0.2n}-x_1} \le 0.02$ and $\frac{x_n-x_{0.98n}}{x_n-x_{0.8n}} \le 0.02$, $\psi_1 = \ln\left(x - x_1\right) - \ln\left(x_n - x\right)$.

And in else situation, we do not transform the random variable.

Step 2:

If $\frac{x_{0.05n}-x_1}{x_{0.5n}-x_{0.05n}} > 1$ and $\frac{x_n-x_{0.95n}}{x_{0.95n}-x_{0.5n}} > 1$, $\psi_2 = \ln\left(cx - cx_{0.5n} + \sqrt{1 + \left(cx - cx_{0.5n}\right)^2}\right)$.

If $\frac{x_{0.05n}-x_1}{x_{0.5n}-x_{0.05n}} > 1$ and $\frac{x_n-x_{0.95n}}{x_{0.95n}-x_{0.5n}} \le 1, \psi_2 = -\ln\left(1 + cx_n - cx\right)$.

If $\frac{x_{0.05n}-x_1}{x_{0.5n}-x_{0.05n}} \le 1$ and $\frac{x_n-x_{0.95n}}{x_{0.95n}-x_{0.5n}} > 1, \psi_2 = \ln\left(1 + cx - cx_1\right)$.

Do the transformations again and again until $\frac{x_{0.05n}-x_1}{x_{0.5n}-x_{0.05n}} \le 1$ and $\frac{x_n-x_{0.95n}}{x_{0.95n}-x_{0.5n}} \le 1$.

where c is the value that makes $Q = \sum\limits_{i=1}^{n} \left[y_i - y\left(x_i\right)\right]^2$ get the minimum.

Step 3:

In all situations, do the transformation $\psi_3 = \frac{5\left(x-x_{0.5n}\right)}{x_{0.95n}-x_{0.05n}}$.

After the three steps, most distributions can be estimated by the spline function. Take note that these transformations focus on the discontinuity of the two ends, but if the discontinuity is in the middle, we should separate the samples to several parts and take the spline regression for each part.

2.3 Spline Function

To define the spline function, basis functions can be set as below:

$$\varphi_{11}(x) = \mathrm{e}^{-x^2}, \varphi_{12}(x) = x\mathrm{e}^{-x^2}, \varphi_{13}(x) = \mathrm{e}^{x}, \varphi_{14}(x) = 1, \varphi_{15}(x) = x,$$

$$\begin{aligned} &\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2, \\ &\varphi_{i4}(x) = x^3, \varphi_{i5}(x) = x^4, \quad i = 2, \dots, 6, \end{aligned} \tag{11}$$

$$\varphi_{71}(x) = \mathrm{e}^{-x^2}, \varphi_{72}(x) = x\mathrm{e}^{-x^2}, \varphi_{73}(x) = \mathrm{e}^{-x}, \varphi_{74}(x) = 1, \varphi_{75}(x) = x.$$

With such predefined basis functions, segments from the middle are quartic spline function, but in the first and last segments, the special basis functions are employed to describe the asymptotic approximation of the distribution function.

The knots $s_2, s_3, \dots, s_u$ are set as the 0.05, 0.23, 0.41, 0.59, 0.77, 0.95 quantile of $x_1, x_2, \dots, x_n$, respectively. If knot s is the a quantile of $x_1, x_2, \dots, x_n$, then $s = x_k$ with k is as the approximate number of $a \cdot n$. with such assumptions above, the first and last segment cover 5% of all sample and the other segments cover 18%, respectively.

Here, we assumed that $u = 7$ segments in all, and $v = 5$ basis functions in each segment, it can easily be obtained that the third order derivative of these functions are still continuous. The value of u may be greater than 7 for the cases of that the distribution function is more complex or the sample size is very large, However, the complexity of function will not increase with the sample size when we take any other parameters, which is quite different from most methods.

As an important parameter, the smooth factor σ will influence the performance of estimation greatly. The proper value of σ to different distribution and different sample size will be discussed in the following sections.

2.4 Adjustment of the Spline Function

With the definition or requirement to CDF, which should not be more than 1, be positive or monotone increasing, and PDF may take negative value, some constrained requirements/conditions should be introduced when we try to estimate CDF or PDF with spline regression, which may lead to much more complex in the calculation. In this paper, a simple method is introduced to resolve such problem by adjusting the spline function after regression.

In most cases, only the first and last segments of the spline function are required be adjusted. In the first segment, the constrained conditions are

$$\begin{pmatrix} \varphi_{11}(s_2) & \varphi_{12}(s_2) & \cdots & \varphi_{1v}(s_2) \\ \varphi'_{11}(s_2) & \varphi'_{12}(s_2) & \cdots & \varphi'_{1v}(s_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{11}^{(v-2)}(s_2) & \varphi_{12}^{(v-2)}(s_2) & \cdots & \varphi_{1v}^{(v-2)}(s_2) \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1v} \end{pmatrix} = \begin{pmatrix} y(s_2) \\ y'(s_2) \\ \vdots \\ y^{(v-2)}(s_2) \end{pmatrix}, \tag{12}$$

with v unknowns in these $v-1$ equations, and only one free variable. For simple, a_{11} is set as the free variable, and Eqs. (12) can be updated as

$$\begin{aligned} &\begin{pmatrix} \varphi_{12}(s_2) & \cdots & \varphi_{1v}(s_2) \\ \varphi'_{12}(s_2) & \cdots & \varphi'_{1v}(s_2) \\ \vdots & \ddots & \vdots \\ \varphi_{12}^{(v-2)}(s_2) & \cdots & \varphi_{1v}^{(v-2)}(s_2) \end{pmatrix} \begin{pmatrix} a_{12} \\ \vdots \\ a_{1v} \end{pmatrix} \\ &= \begin{pmatrix} y(s_2) \\ y'(s_2) \\ \vdots \\ y^{(v-2)}(s_2) \end{pmatrix} - a_{11} \begin{pmatrix} \varphi_{11}(s_2) \\ \varphi'_{11}(s_2) \\ \vdots \\ \varphi_{11}^{(v-2)}(s_2) \end{pmatrix}. \end{aligned} \tag{13}$$

Values of a_{1j}, $j = 2, \ldots, v$ can be derived based on the initial set a_{11}, different preset value of a_{11} will be repeated until we got the reasonable estimated CDF and PDF based on all samples calculated by the spline function. The last segment is adjusted in the same approach. In some cases of that the reasonable result is not available via one free value, the constrained conditions should be reduced to have two free values introduced.

The algorithm to estimate of probability distribution with spline regression model is summarized as Algorithm 1.

```
Algorithm 1 Estimation to probability distribution with spline regression model
Step 1. Transform the ordered samples: $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}$ with step in section 2.2;
Step 2. Set basis functions as basis section 2.3 as the spline function.
Step 3. With equation (3) – (9), the spline regression is built and to have probability distribution
        estimated.
```

2.5 Method Evaluation and Comparison

For these 40 distributions (Table 1) with different types or parameters, the characteristics of these most widely used statistics has been considered to evaluate the performance of estimated CDF or PDF. However, with the increase of sample size, integrated square error (ISE), mean absolute error (MAE) and mean square error (MSE) are not valid statistics to evaluate the estimate PDF. For the example of distribution $F(x) = \sqrt[3]{x}\,\text{and}\,f(x) = \frac{1}{3}x^{-\frac{2}{3}}, x \in (0, 1)$, sample data $x_i = (i/n)^3 i = 1, \ldots, n-1$, and estimated PDF

$$y'(x) = \begin{cases} \frac{1.01}{3}x^{-\frac{2}{3}}, x \in \left(0, \frac{1}{8}\right] \\ \frac{0.99}{3}x^{-\frac{2}{3}}, x \in \left(\frac{1}{8}, 1\right) \end{cases} \text{when error is } 1\%,$$

$$\text{integrated absolute error (IAE)} = \int_{x_1}^{x_n} \left|y'(x) - f(x)\right| \mathrm{d}x = 0.01 \cdot \frac{n-2}{n},$$

integrated square error (ISE)

$$= \int_{x_1}^{x_n} \left(y'(x) - f(x)\right)^2 \mathrm{d}x = \frac{0.01^2}{3}\left(n - \frac{n}{n-1}\right),$$

mean absolute error (MAE)

$$= \frac{1}{n}\sum_{i=1}^{n} \left|y'^{(x_i)} - f(x_i)\right| = \frac{0.01}{3} \cdot \frac{n^2}{n-1} \sum_{i=1}^{n-1} \frac{1}{i^2},$$

mean square error (MSE)

$$= \frac{1}{n}\sum_{i=1}^{n} \left(y'^{(x_i)} - f(x_i)\right)^2 = \frac{0.01^2}{9} \cdot \frac{n^4}{n-1} \sum_{i=1}^{n-1} \frac{1}{i^4},$$

then $\text{IAE} \to 0.01, \text{ISE} \to +\infty, \text{MAE} \to +\infty, \text{MSE} \to +\infty$ when $n \to \infty$. Integrated absolute error (IAE) is used as the statistics to evaluate the estimated PDF. Similarly, IAE and ISE are not convergent with the increase of sample data, root mean square error $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(x_i) - F(x_i)\right)^2}$ is to evaluate the estimated CDF.

And

$$\text{IAE} = \int_{x_1}^{x_n} \left|y'(x) - f(x)\right| \mathrm{d}x \le \int_{x_1}^{x_n} \left(y'(x) + f(x)\right) \mathrm{d}x$$

$$= y(x_n) - y(x_1) + F(x_n) - F(x_1) \le 2.$$

For example, if random variable X follows the distribution:

$$F(x) = \sqrt[3]{x}\,\text{and}\,f(x) = \frac{1}{3}x^{-\frac{2}{3}}, x \in (0, 1),$$

is a set of samples taken from random variable X.

In the case of estimated error is in 1%, and the estimated PDF is $y'(x) = \begin{cases} \frac{1.01}{3}x^{-\frac{2}{3}}, x \in \left(0, \frac{1}{8}\right] \\ \frac{0.99}{3}x^{-\frac{2}{3}}, x \in \left(\frac{1}{8}, 1\right) \end{cases}$, for each different kind of statistics to evaluate the performance of PDF estimation as below.

Table 1 40 Distributions used in the method evaluation and comparison

No.	Distribution	Density function
1	Beta (2,2)	$f(x) = 6x(1-x), x \in (0,1)$
2	Beta (2,1)	$f(x) = 2x, x \in (0,1)$
3	Beta (0.5,1)	$f(x) = \frac{1}{2\sqrt{x}}, x \in (0,1)$
4	Beta (0.5,0.5)	$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, x \in (0,1)$
5	Beta (1/3,2/3)	$f(x) = \frac{\sqrt{3}}{2\pi} x^{-\frac{2}{3}} (1-x)^{-\frac{1}{3}}, x \in (0,1)$
6	Beta (0.2,5)	$f(x) = \frac{924}{3125} x^{-\frac{4}{5}} (1-x)^4, x \in (0,1)$
7	Beta (5,10)	$f(x) = 10010x^4(1-x)^9, x \in (0,1)$
8	Exponential (1)	$f(x) = \mathrm{e}^{-x}, x \in [0,+\infty)$
9	Exponential (0.1)	$f(x) = 10\mathrm{e}^{-10x}, x \in [0,+\infty)$
10	Exponential (10)	$f(x) = \frac{1}{10}\mathrm{e}^{-\frac{x}{10}}, x \in [0,+\infty)$
11	Extreme value (0,1)	$f(x) = \mathrm{e}^{x-e^x}$
12	Gamma (2,1)	$f(x) = x\mathrm{e}^{-x}, x \in (0,+\infty)$
13	Gamma (0.5,0.2)	$f(x) = \frac{\sqrt{5\pi}}{\pi} x^{-\frac{1}{2}} \mathrm{e}^{-5x}, x \in (0,+\infty)$
14	Gamma (5,0.2)	$f(x) = \frac{3125}{24} x^4 \mathrm{e}^{-5x}, x \in (0,+\infty)$
15	Generalized extreme value (0.2,5,0)	$f(x) = \frac{1}{5}\left(1+\frac{x}{25}\right)^{-5} \exp\left(-\left(1+\frac{x}{25}\right)^{-4}\right), x \in (-25,+\infty)$
16	Generalized extreme value (− 2,5,0)	$f(x) = \frac{1}{5}(1-0.4x)^{-\frac{1}{2}} \exp\left(-(1-0.4x)^{\frac{1}{2}}\right), x \in \left(-\infty, \frac{5}{2}\right)$
17	Generalized extreme value (− 5,0.2,0)	$f(x) = 5(1-25x)^{-\frac{4}{5}} \exp\left(-(1-25x)^{\frac{1}{5}}\right), x \in \left(-\infty, \frac{1}{25}\right)$
18	Generalized extreme value (1,1,0)	$f(x) = \frac{1}{(1+x)^2} \exp\left(-\frac{1}{1+x}\right), x \in (-1,+\infty)$
19	Generalized extreme value (2,0.2,0)	$f(x) = 5(1+10x)^{-\frac{3}{2}} \exp\left(-(1+10x)^{-\frac{1}{2}}\right), x \in \left(-\frac{1}{10},+\infty\right)$
20	Generalized extreme value (5,5,0)	$f(x) = \frac{1}{5}(1+x)^{-\frac{4}{5}} \exp\left(-(1+x)^{-\frac{1}{5}}\right), x \in (-1,+\infty)$
21	Generalized Pareto (1,1,0)	$f(x) = \frac{1}{(1+x)^2}, x \in (0,+\infty)$
22	Generalized Pareto (0.2,5,0)	$f(x) = \frac{1}{5}\left(1+\frac{x}{25}\right)^{-6}, x \in (0,+\infty)$
23	Generalized Pareto (− 0.2,5,0)	$f(x) = \frac{1}{5}\left(1-\frac{x}{25}\right)^{4}, x \in (0,25)$
24	Generalized Pareto (− 5,0.2,0)	$f(x) = 5(1-25x)^{-\frac{4}{5}}, x \in \left(0, \frac{1}{25}\right)$
25	Lognormal (0,1)	$f(x) = \frac{1}{\sqrt{2\pi}x}\mathrm{e}^{-\frac{(\ln x)^2}{2}},$
26	Lognormal (2,5)	$f(x) = \frac{1}{5\sqrt{2\pi}x}\mathrm{e}^{-\frac{(\ln x-2)^2}{50}}, x \in (0,+\infty)$
27	Lognormal (5,0.2)	$f(x) = \frac{5}{\sqrt{2\pi}x}\mathrm{e}^{-\frac{25(\ln x-5)^2}{2}}, x \in (0,+\infty)$
28	Normal (0,1)	$f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$
29	Rayleigh (1)	$f(x) = x\mathrm{e}^{-\frac{x^2}{2}}, x \in (0,+\infty)$
30	Weibull (2,3)	$f(x) = \frac{3x^2}{8}\mathrm{e}^{-\frac{x^3}{8}}, x \in (0,+\infty)$
31	Logistic (0,1)	$f(x) = \frac{e^x}{(1+e^x)^2}$
32	Uniform (0,1)	$f(x) = 1, x \in [0,1]$
33	χ^2 (1)	$f(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \mathrm{e}^{-\frac{x}{2}}, x \in (0,+\infty)$
34	χ^2 (3)	$f(x) = \frac{1}{\sqrt{2\pi}} x^{\frac{3}{2}} \mathrm{e}^{-\frac{x}{2}}, x \in (0,+\infty)$

Table 1 (continued)

No.	Distribution	Density function
35	$t\,(1)$	$f(x) = \frac{1}{\pi\left(1+x^2\right)}$
36	$t\,(3)$	$f(x) = \frac{2}{\sqrt{3}\pi}\left(1 + \frac{x^2}{3}\right)^{-2}$
37	$t\,(6)$	$f(x) = \frac{15}{16\sqrt{6}}\left(1 + \frac{x^2}{6}\right)^{-\frac{7}{2}}$
38	$F\,(4{,}4)$	$f(x) = \frac{6x}{(1+x)^4}, x \in (0, +\infty)$
39	$F\,(4{,}10)$	$f(x) = \frac{24x}{5(1+0.4x)^7}, x \in (0, +\infty)$
40	$F\,(10{,}4)$	$f(x) = \frac{24x^4}{5(0.4+x)^7}, x \in (0, +\infty)$

$$\text{integrated absolute error(IAE)} = \int_{x_1}^{x_n} \left|y'(x) - f(x)\right| \mathrm{d}x = 0.01 \cdot \frac{n-2}{n},$$

integrated square error (ISE)

$$= \int_{x_1}^{x_n} \left(y'(x) - f(x)\right)^2 \mathrm{d}x = \frac{0.01^2}{3}\left(n - \frac{n}{n-1}\right),$$

mean absolute error (MAE)

$$= \frac{1}{n}\sum_{i=1}^{n}\left|y'^{(x_i)} - f\left(x_i\right)\right| = \frac{0.01}{3} \cdot \frac{n^2}{n-1}\sum_{i=1}^{n-1}\frac{1}{i^2},$$

mean square error (MSE)

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y'^{(x_i)} - f\left(x_i\right)\right)^2 = \frac{0.01^2}{9} \cdot \frac{n^4}{n-1}\sum_{i=1}^{n-1}\frac{1}{i^4}.$$

When $\mathrm{IAE} \to 0.01, \mathrm{ISE} \to +\infty, \mathrm{MAE} \to +\infty, \mathrm{MSE} \to +\infty$, only IAE is a valid statistics to evaluate the estimation. Similarly, IAE and ISE are not suitable to evaluate the performance of estimated CDF due to that they are not always convergent.

3 Result

3.1 Basis Functions

For distributions normal (0,1), exponential (1) and Rayleigh (1), 1000 random samples were generated with Monte-Carlo simulation. Six different sets of basis functions as below were employed in the estimation of CDF and PDF.

$E1$:

$$\varphi_{11}(x) = \mathrm{e}^{-x^2}, \varphi_{12}(x) = x\mathrm{e}^{-x^2}, \varphi_{13}(x) = \mathrm{e}^{x}, \varphi_{14}(x) = 1, \varphi_{15}(x) = x,$$

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2,$$
$$\varphi_{i4}(x) = x^3, \varphi_{i5}(x) = x^4 \quad i = 2, \ldots, 6,$$

$$\varphi_{71}(x) = \mathrm{e}^{-x^2}, \varphi_{72}(x) = x\mathrm{e}^{-x^2}, \varphi_{73}(x) = \mathrm{e}^{-x}, \varphi_{74}(x) = 1, \varphi_{75}(x) = x.$$

$E2$:

$$\varphi_{11}(x) = \mathrm{e}^{-x^2}, \varphi_{12}(x) = \mathrm{e}^{x}, \varphi_{13}(x) = 1, \varphi_{14}(x) = x, \varphi_{15}(x) = x^2,$$

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2,$$
$$\varphi_{i4}(x) = x^3, \varphi_{i5}(x) = x^4 \quad i = 2, \ldots, 6,$$

$$\varphi_{71}(x) = \mathrm{e}^{-x^2}, \varphi_{72}(x) = \mathrm{e}^{-x}, \varphi_{73}(x) = 1, \varphi_{74}(x) = x, \varphi_{75}(x) = x^2.$$

$E3$:

$$\varphi_{11}(x) = \mathrm{e}^{x}, \varphi_{12}(x) = 1, \varphi_{13}(x) = x, \varphi_{14}(x) = x^2, \varphi_{15}(x) = x^3,$$

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2,$$
$$\varphi_{i4}(x) = x^3, \varphi_{i5}(x) = x^4 \quad i = 2, \ldots, 6,$$

$$\varphi_{71}(x) = \mathrm{e}^{-x}, \varphi_{72}(x) = 1, \varphi_{73}(x) = x, \varphi_{74}(x) = x^2, \varphi_{75}(x) = x^3.$$

$E4$:

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2,$$
$$\varphi_{i4}(x) = x^3, \varphi_{i5}(x) = x^4 \quad i = 1, \ldots, 7.$$

$E5$:

$$\varphi_{11}(x) = \mathrm{e}^{-x^2}, \varphi_{12}(x) = x\mathrm{e}^{-x^2}, \varphi_{13}(x) = \mathrm{e}^{x},$$
$$\varphi_{14}(x) = x\mathrm{e}^{x}, \varphi_{15}(x) = 1, \varphi_{16}(x) = x,$$

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2, \varphi_{i4}(x) = x^3,$$
$$\varphi_{15}(x) = x^4, \varphi_{i6}(x) = x^5 \quad i = 2, \ldots, 6,$$

$$\varphi_{71}(x) = \mathrm{e}^{-x^2}, \varphi_{72}(x) = x\mathrm{e}^{-x^2}, \varphi_{73}(x) = \mathrm{e}^{-x},$$
$$\varphi_{74}(x) = x\mathrm{e}^{-x}, \varphi_{75}(x) = 1, \varphi_{76}(x) = x.$$

$E6$:

$$\varphi_{i1}(x) = 1, \varphi_{i2}(x) = x, \varphi_{i3}(x) = x^2, \varphi_{i4}(x) = x^3,$$
$$\varphi_{i5}(x) = x^4, \varphi_{i6}(x) = x^5 \quad i = 1, \ldots, 7.$$

With results in Table 2, we can see the $\mathrm{RMSE}_{\mathrm{CDF}}$ is not sensitive to the basis functions but $\mathrm{IAE}_{\mathrm{PDF}}$ is influenced by the basis functions significantly. Based on Fig. 1, we can also find that the two ends of the distribution are usually hard to be estimated but basis function of $E1$ shows the best estimation and successfully describes the asymptotic approximation, which has significant advantage over pure polynomial spline function ($E4$ and $E6$). Then we use $E1$ as the basis functions, which has been mentioned in Sect. 2.3.

3.2 Smooth Factor

With Monte Carlo simulation, IAEs for each batch of generated random samples were repeated for 50 times and averaged to evaluate the estimated PDF. In Fig. 2a, with the increase of smooth factor σ, the averaged IAEs for most distributions decreases at first and increases after getting the minimum, but the averaged IAE to uniform distribution is in decreases with the increase of smooth factor σ. The minimum averaged IAE is for different values to smooth factor σ. and when $\sigma = 2$, the averaged IAE to most distributions is in the nearby area of minimum averaged IAE. In Fig. 2b, σ corresponding to the minimum averaged IAE is identical with different sample sizes, and with the increase of sample size, the averaged IAE is in smaller with the increase number of simulated sample. The minimum value of averaged IAEs are in the nearby area of $\sigma = 2$. Then $\sigma = 2$ is chosen as the optimal smooth factor for each different distribution and different sample size.

3.3 Evaluation for Well-Proportioned Samples

In the extreme case of well-proportioned samples $x_1, x_2, \ldots, x_n \left(x_1 \le x_2 \le \cdots \le x_n\right)$, which means $F\left(x_i\right) = y_i = i/(n+1)$ can be well estimated. With Monte Carlo simulation, all these 40 distributions from Sect. 2.5 were evaluated using this well-proportioned sample. As Fig. 3, RMSEs to these estimated CDFs are all smaller than 0.00016 and IAEs to these estimated PDFs are all smaller than 0.008, which means that the proposed method estimated most distributions well.

3.4 Evaluation for Random Samples

With Monte Carlo simulation, for each distribution from Sect. 2.5, RMSE and IAE for each batch of generated random samples were repeated for 50 times and averaged to evaluate the estimated CDF and PDF. Each batch random samples were constructed in the steps as below:

(1) Sort the sampled n random samples from standard uniform distribution $U(0,1)$.
(2) Calculate $F^{-1}(x)$ (inverse function of the distribution function) with these n samples, as the random samples for each distribution.

As the CDF is follow the standard unit distribution $U(0,1)$, for each set of random samples,

$$\mathrm{RMSE}_{\mathrm{rand}} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{i}{n+1} - F\left(x_i\right)\right)^2},$$

is to evaluate the deviation from distribution of these samples, which can also be seen as the error of empirical distribution function with $i/(n+1)$ as the estimation of $F\left(x_i\right)$. In Fig. 4, $\mathrm{RMSE}_{\mathrm{rand}}$ for our proposed method is almost always smaller than $\mathrm{RMSE}_{\mathrm{rand}}$ for any distribution, which indicates that our method is superior to empirical distribution function.

In PDF estimation, compared to kernel density estimation in Fig. 5 and Table 3, our proposed method is superior in the estimation of PDF for any distribution. Most of the IAE is in the range of 0.05–0.06 on average, while kernel density estimation is generated larger error than the proposed method. In Fig. 6, both normal distribution and Rayleigh distribution can be well estimated with both kernel density estimation with normal kernel function and spline regression, but spline regression is much more accuracy than kernel density estimation with normal kernel function in the estimation of PDF for all of these distributions.

Table 2 The evaluation result of different basis functions

Basis functions	Normal (0,1)		Exponential (1)		Rayleigh (1)	
	$\mathrm{RMSE}_{\mathrm{CDF}}$	$\mathrm{IAE}_{\mathrm{PDF}}$	$\mathrm{RMSE}_{\mathrm{CDF}}$	$\mathrm{IAE}_{\mathrm{PDF}}$	$\mathrm{RMSE}_{\mathrm{CDF}}$	$\mathrm{IAE}_{\mathrm{PDF}}$
$E1$	0.0116	0.0466	0.0119	0.0507	0.0117	0.0431
$E2$	0.0116	0.0808	0.0119	0.0553	0.0118	0.0444
$E3$	0.0116	0.0729	0.0119	0.0504	0.0118	0.0501
$E4$	0.0116	0.0816	0.0119	0.0503	0.0118	0.0495
$E5$	0.0116	0.0702	0.0119	0.0506	0.0117	0.0461
$E6$	0.0116	0.1196	0.0119	0.0572	0.0117	0.0703

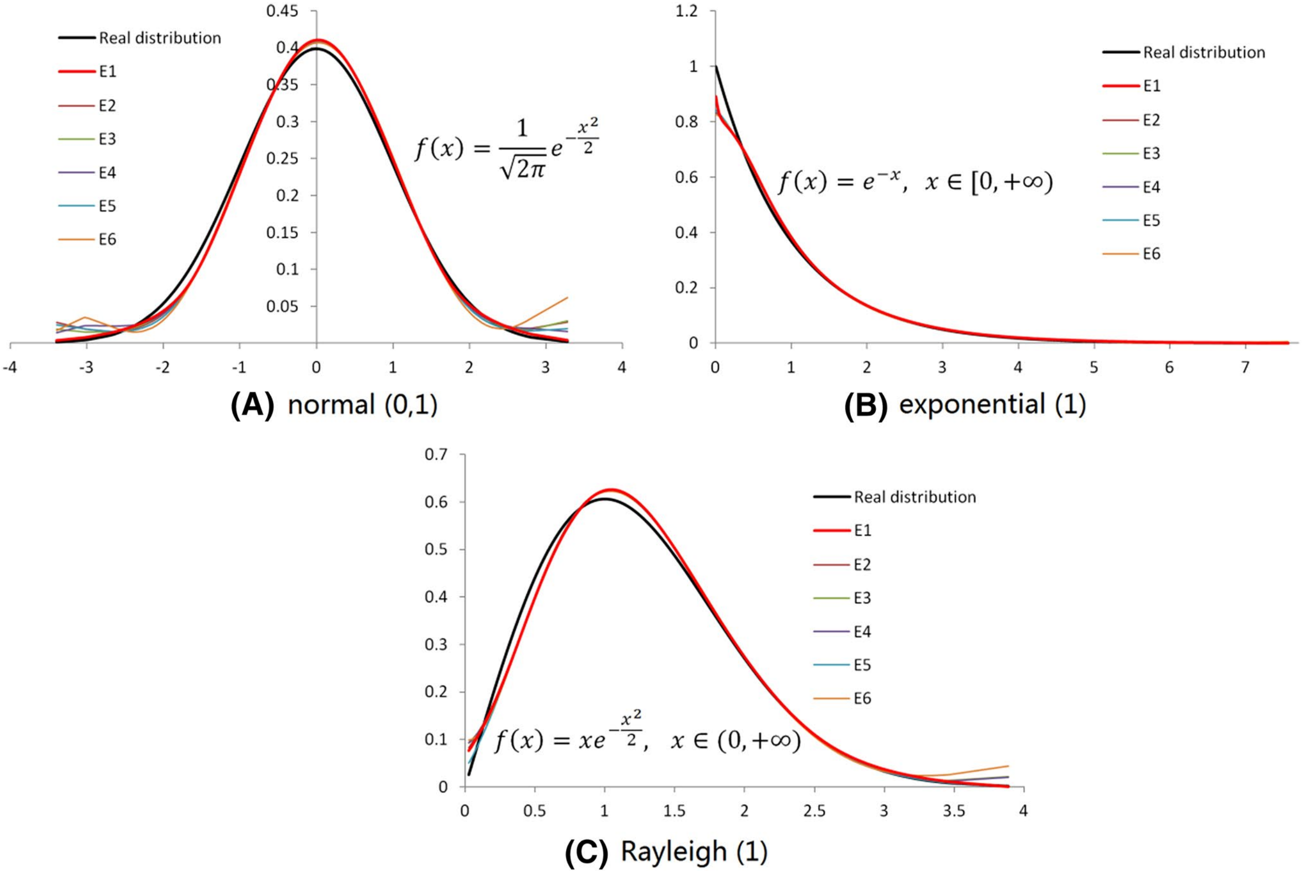


Fig. 1 Illustrations of the density functions estimated by different basis functions. **a** Normal distribution. **b** Exponential distribution. **c** Rayleigh distribution

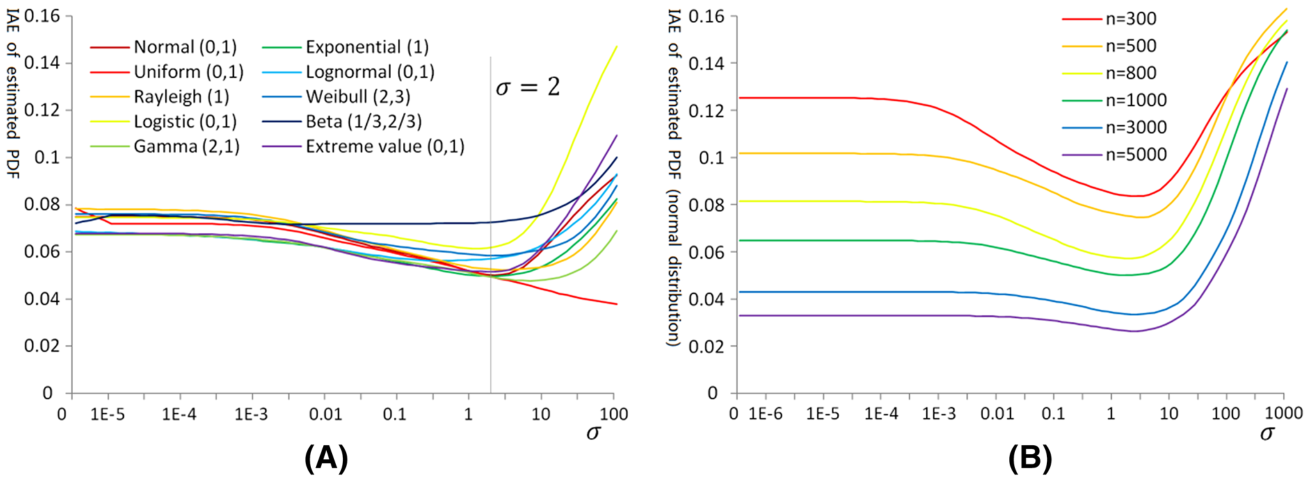


Fig. 2 The influence of smooth factor σ to the estimated error of PDF. **a** With different distributions. **b** With different sample size

3.5 Large Sample Size

With the much large sample size, CDF and PDF are estimated by a subset of the sample. With Monte Carlo simulation, the similar accuracy is for both our method for the subset sample and the full sample. In the simulated samples, the subset sample is obtained via the every 100 data for this ascending sorted sample when the sample size is

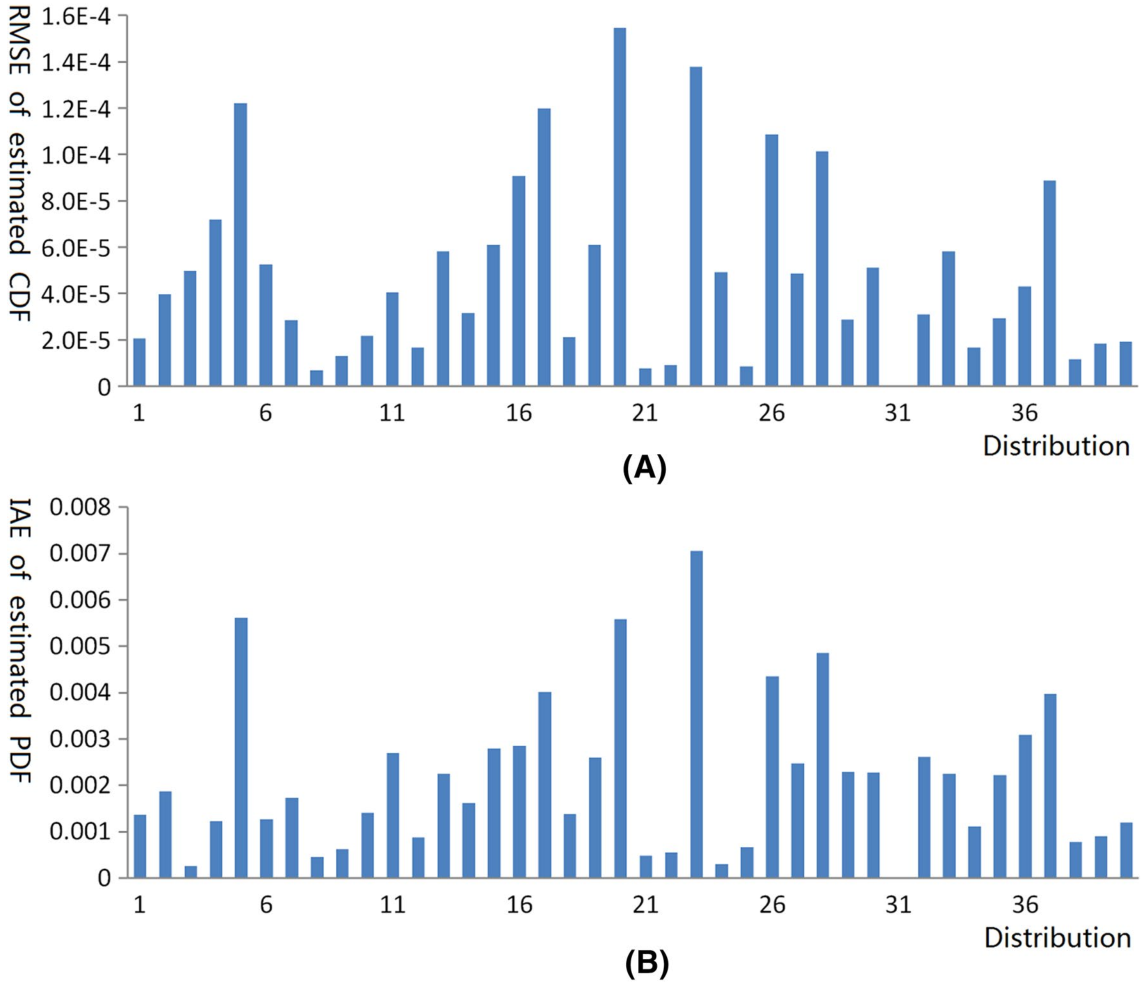


Fig. 3 RMSE of estimated CDF and IAE of estimated PDF to well-proportioned samples

$n = 100{,}000$, then the subset sample of these 1000 observations are used to estimate CDF and PDF. With Table 4, the RMSE to estimated CDF and IAE to estimated PDF are quite similar between the full sample and the subset samples.

4 Discussion

In this section, the PDF estimation to high dimensional random variables and the application in classification and regression models will be discussed.

4.1 Probability Distribution of *n* Dimensional Random Variables

It is almost impossible to estimate the joint probability distribution of n dimensional random variables by limited number of samples because of the curse of dimensionality, but the problem can be simplified as the linear correlations of every variables which will be a rough but quite practical approach in the estimation.

To simplify the problem, the following assumption is to be hold:

If random variables $Y_1, Y_2, \ldots, Y_n$ follow normal distribution, n dimensional random variable $(Y_1, Y_2, \ldots, Y_n)$ follows n dimensional joint normal distribution approximately.

This approximation uses normal distribution as a bridge to construct high dimensional probability distribution. Then, for any n dimensional random variable $(X_1, X_2, \ldots, X_n)$, set the marginal distribution functions as $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ and define

$$\hat{X}_i = \Phi^{-1}\left(F_i(X_i)\right) \quad i = 1, \ldots, n.$$

Then $P(\hat{X}_i \le \hat{x}_i) = P\left(\Phi^{-1}\left(F_i(X_i)\right) \le \hat{x}_i\right) = P\left(X_i \le F_i^{-1}\left(\Phi(\hat{x}_i)\right)\right) = F_i\left(F_i^{-1}\left(\Phi(\hat{x}_i)\right)\right) = \Phi(\hat{x}_i)$, where $\Phi(x)$ is the distribution function of standard normal distribution. And we can see $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ follow normal distribution.

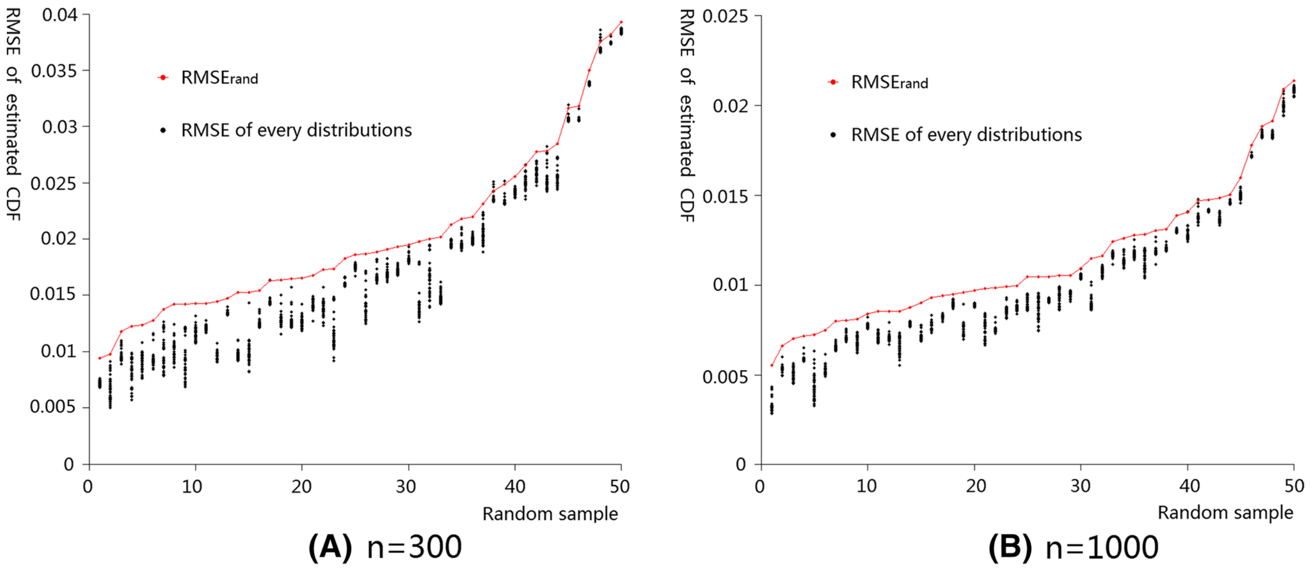


Fig. 4 In each set of random samples, red dots are the $\mathrm{RMSE}_{\mathrm{rand}}$ for each set of random numbers generated by step (1). Corresponding series of random samples to each set of random sample from step (1) is generated for each different distribution by step (2). RMSE to each estimated CDF was represented by black dots. **a** The sample size is 300. **b** The sample size is 1000

Based on the previous assumption, the joint distribution function of n dimensional random variable $(X_1, X_2, \ldots, X_n)$ can be estimated as

$$
\begin{aligned}
F(x_1, x_2, \ldots, x_n) &= P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) \\
&= P\left(F_1^{-1}(\boldsymbol{\Phi}(\hat{X}_1)) \le x_1, F_2^{-1}(\boldsymbol{\Phi}(\hat{X}_2)) \le x_2, \ldots, F_n^{-1}(\boldsymbol{\Phi}(\hat{X}_n)) \le x_n\right) \\
&= P\left(\hat{X}_1 \le \boldsymbol{\Phi}^{-1}(F_1(x_1)), \hat{X}_2 \le \boldsymbol{\Phi}^{-1}(F_2(x_2)), \ldots, \hat{X}_n \le \boldsymbol{\Phi}^{-1}(F_n(x_n))\right) \approx \boldsymbol{\Phi}_n(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)
\end{aligned}
$$

And the joint density function is

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n) \\
&\approx \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} \boldsymbol{\Phi}_n\left(\boldsymbol{\Phi}^{-1}(F_1(x_1)), \boldsymbol{\Phi}^{-1}(F_2(x_2)), \ldots, \boldsymbol{\Phi}^{-1}(F_n(x_n))\right) \\
&= \varphi_n\left(\boldsymbol{\Phi}^{-1}(F_1(x_1)), \boldsymbol{\Phi}^{-1}(F_2(x_2)), \ldots, \boldsymbol{\Phi}^{-1}(F_n(x_n))\right) \cdot \prod_{i=1}^{n} \frac{f_i(x_i)}{\varphi\left(\boldsymbol{\Phi}^{-1}(F_i(x_i))\right)} \\
&= |R|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\hat{x}(R^{-1} - E)\hat{x}^T\right) \cdot \prod_{i=1}^{n} f_i(x_i) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\ln\lambda_i + \frac{\beta_i^2}{\lambda_i} - \beta_i^2 - 2\ln f_i(x_i)\right)\right]
\end{aligned}
$$

where $\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, $R = (\rho_{ij})$ is the correlation coefficient matrix of $(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)$, λ_i and α_i is the eigenvalues and eigenvectors of R, respectively, and $(\beta_1, \beta_2, \ldots, \beta_n) = \hat{x}(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

$\varphi_n(x) = (2\pi)^{-\frac{1}{2}}|R|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}xR^{-1}x^{\mathrm{T}}\right), x = (x_1, x_2, \ldots, x_n)$ is the density function of n dimensional standard normal distribution, and $\boldsymbol{\Phi}_n(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \varphi_n(t_1, \ldots, t_n)\mathrm{d}t_1 \cdots \mathrm{d}t_n$ is the distribution function.

4.2 Application in Bayesian Classification

For the sample with n features $(x_1, x_2, \ldots, x_n)$, based on Bayes' theorem, the probability that it belongs to class C_k is

$$
P(C_k | x_1, x_2, \ldots, x_n) = \frac{P(C_k)P(x_1, x_2, \ldots, x_n | C_k)}{P(x_1, x_2, \ldots, x_n)},
$$

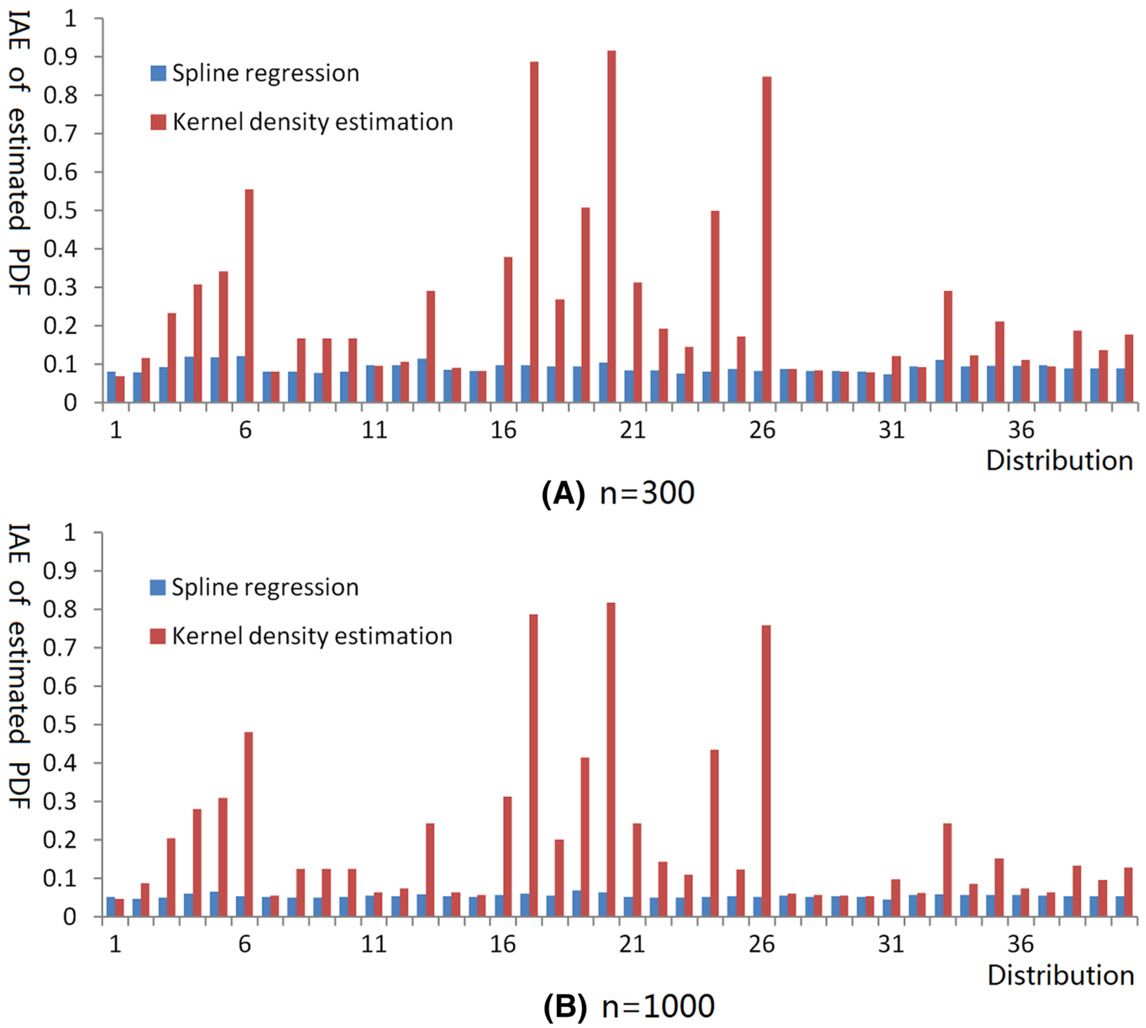


Fig. 5 Comparision of spline regression and kernel density function with **a** the sample size is 300. **b** the sample size is 1000

where $P(C_k)$ is the probability that any sample belongs to class C_k, $P(x_1, x_2, \ldots, x_n | C_k)$ is prior and $P(C_k | x_1, x_2, \ldots, x_n)$ is posterior.

With the assumption of that, every features are independent of each other for a given class label, which is independence of conditional probability. Then

$$P(C_k | x_1, x_2, \ldots, x_n) = \frac{P(C_k)}{P(x_1, x_2, \ldots, x_n)} \prod_{i=1}^{n} P(x_i | C_k)$$

Take note that $P(x_1, x_2, \ldots, x_n)$ is independent of C_k, so we can get the predictive classification of $(x_1, x_2, \ldots, x_n)$

$$\hat{y} = \underset{k \in 1,2,\ldots,K}{\arg\max}\, P(C_k) \prod_{i=1}^{n} P(x_i | C_k).$$

This method is naïve Bayes classifier, a basic algorithm in machine learning and shows implausible efficacy in many complex real-world situations [27–29].

But release such strong assumption and based on the estimation of the density function of n dimensional random variables, the predictive classification can be calculated straight forwardly:

$$\hat{y} = \underset{k \in 1,2,\ldots,K}{\arg\max}\, P(C_k) P(x_1, x_2, \ldots, x_n | C_k).$$

With such update, we not only include the correlation of features into the model prediction, but also greatly extend the application of Bayesian classification.

Table 3 The evaluation result of spline regression and kernel density estimation by random samples

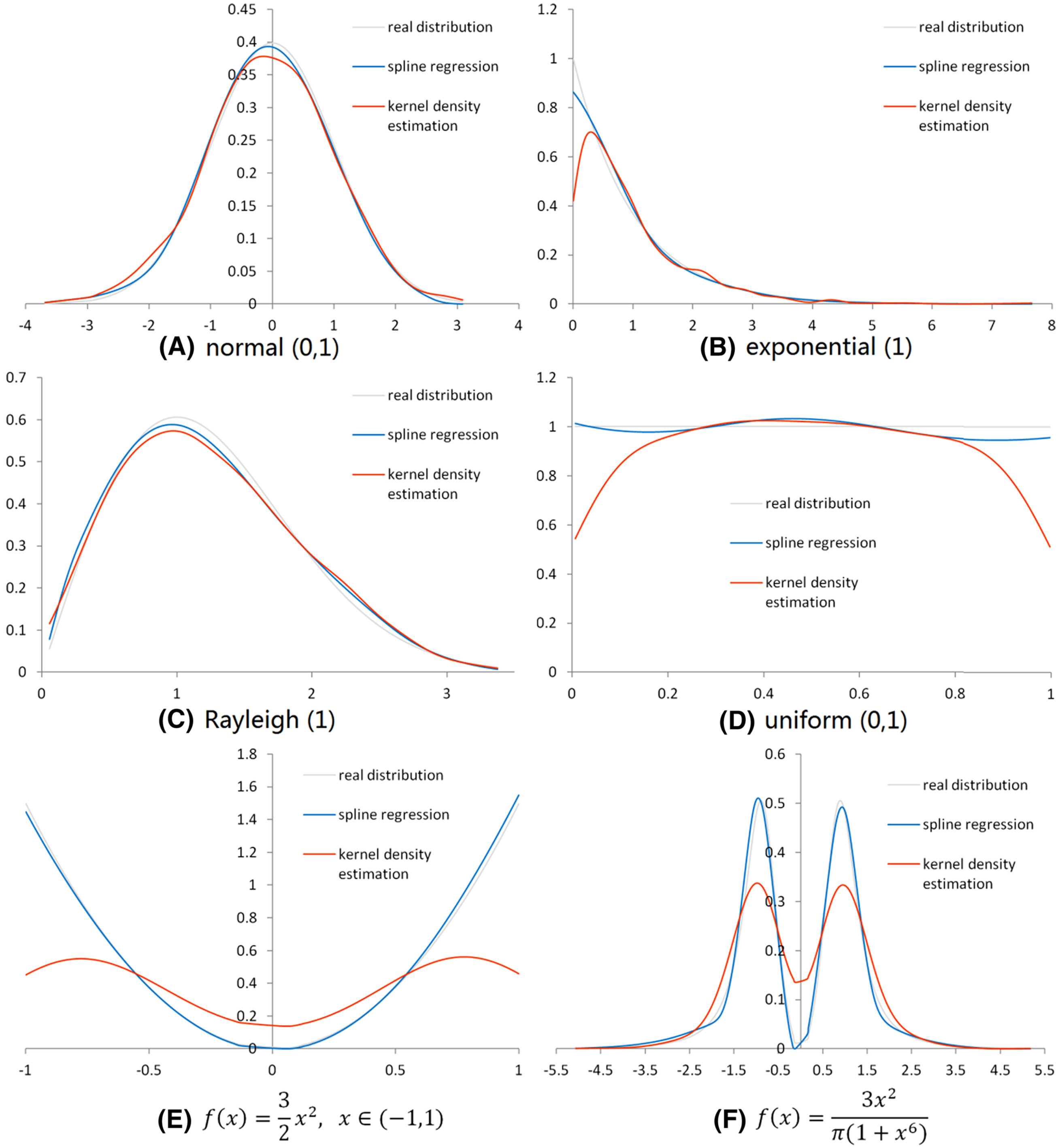
Distribution	RMSE of CDF with spline regression		IAE of PDF			
			Spline regression		Kernel density estimation	
	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
2	0.0175	0.0100	0.0801	0.0477	0.1166	0.0884
3	0.0176	0.0100	0.0938	0.0512	0.2346	0.2052
4	0.0184	0.0103	0.1205	0.0615	0.3077	0.2806
5	0.0183	0.0103	0.1183	0.0657	0.3419	0.3105
6	0.0181	0.0100	0.1219	0.0545	0.5562	0.4813
7	0.0170	0.0099	0.0806	0.0521	0.0817	0.0566
8	0.0173	0.0101	0.0810	0.0510	0.1682	0.1257
9	0.0172	0.0100	0.0772	0.0503	0.1682	0.1257
10	0.0173	0.0100	0.0820	0.0519	0.1682	0.1257
11	0.0175	0.0099	0.0990	0.0557	0.0970	0.0639
12	0.0175	0.0100	0.0988	0.0549	0.1064	0.0748
13	0.0179	0.0101	0.1153	0.0597	0.2922	0.2445
14	0.0171	0.0099	0.0856	0.0534	0.0907	0.0637
15	0.0171	0.0099	0.0826	0.0524	0.0834	0.0573
16	0.0175	0.0100	0.0985	0.0583	0.3797	0.3132
17	0.0175	0.0101	0.0985	0.0608	0.8881	0.7874
18	0.0175	0.0100	0.0946	0.0556	0.2693	0.2024
19	0.0176	0.0103	0.0956	0.0700	0.5089	0.4144
20	0.0178	0.0101	0.1050	0.0643	0.9175	0.8182
21	0.0174	0.0101	0.0841	0.0529	0.3142	0.2432
22	0.0175	0.0100	0.0852	0.0515	0.1930	0.1441
23	0.0171	0.0100	0.0766	0.0506	0.1459	0.1102
24	0.0174	0.0101	0.0818	0.0528	0.4998	0.4354
25	0.0173	0.0100	0.0880	0.0542	0.1726	0.1239
26	0.0172	0.0099	0.0832	0.0524	0.8488	0.7589
27	0.0173	0.0100	0.0872	0.0560	0.0876	0.0607
28	0.0172	0.0098	0.0832	0.0521	0.0840	0.0568
29	0.0170	0.0099	0.0829	0.0539	0.0814	0.0565
30	0.0171	0.0099	0.0814	0.0519	0.0801	0.0549
31	0.0174	0.0101	0.0749	0.0458	0.1218	0.0977
32	0.0178	0.0102	0.0947	0.0570	0.0929	0.0622
33	0.0178	0.0101	0.1120	0.0594	0.2922	0.2445
34	0.0174	0.0100	0.0951	0.0568	0.1230	0.0870
35	0.0180	0.0103	0.0968	0.0577	0.2114	0.1522
36	0.0179	0.0102	0.0967	0.0575	0.1123	0.0748
37	0.0179	0.0101	0.0979	0.0567	0.0954	0.0640
38	0.0174	0.0100	0.0903	0.0546	0.1882	0.1344
39	0.0173	0.0099	0.0895	0.0546	0.1377	0.0969
40	0.0173	0.0100	0.0892	0.0545	0.1778	0.1282

4.3 Application in Maximum Likelihood Regression

With the proposed approach in the CDF and PDF estimation, maximum likelihood estimation can be extended from parameter estimation [30] to regression models.

The maximum likelihood function to sample $\left(x_1, x_2, \dots, x_n\right)$ is $L(y) = P\left(Y = y|x_1, x_2, \dots, x_n\right)$ with $L(y)$ get the maximum value at the point y.

Then



(A) normal (0,1)

(B) exponential (1)

(C) Rayleigh (1)

(D) uniform (0,1)

(E) $f(x) = \frac{3}{2}x^2, \quad x \in (-1,1)$

(F) $f(x) = \frac{3x^2}{\pi(1 + x^6)}$

Fig. 6 Illustrations of the density functions estimated by spline regression and kernel density function with different distributions

Table 4 The evaluation result of the subsets of large samples

Sample	Size	RMSE of CDF	IAE of PDF
1	100,000	0.000957	0.0109
2	10,000	0.000932	0.0116
3	1000	0.000920	0.0111

$$
\begin{aligned}
L(y) &= \frac{P\left(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, Y = y\right)}{P\left(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\right)} \\
&= \frac{\begin{vmatrix} R & r^{\mathrm{T}} \\ r & 1 \end{vmatrix}^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\hat{x}, \hat{y})\left(\begin{pmatrix} R & r^{\mathrm{T}} \\ r & 1 \end{pmatrix}^{-1} - E\right)(\hat{x}, \hat{y})^{\mathrm{T}}\right) \cdot f_Y(y) \cdot \prod_{i=1}^{n} f_n\left(x_i\right)}{|R|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\hat{x}, \hat{y})\left(\begin{pmatrix} R^{-1} & O \\ O & 1 \end{pmatrix} - E\right)(\hat{x}, \hat{y})^{\mathrm{T}}\right) \cdot \prod_{i=1}^{n} f_n\left(x_i\right)} \\
&= \frac{f_Y(y)}{\sqrt{1 - rR^{-1}r^{T}}} \exp\left(-\frac{1}{2}(\hat{x}, \hat{y})\left(\begin{pmatrix} R & r^{\mathrm{T}} \\ r & 1 \end{pmatrix}^{-1} - \begin{pmatrix} R^{-1} & O \\ O & 1 \end{pmatrix}\right)(\hat{x}, \hat{y})^{\mathrm{T}}\right),
\end{aligned}
$$

where $F_Y(y)$ and $f_Y(y)$ is the distribution and density function of Y, respectively, $\hat{x} = \left(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\right)$, $\hat{y} = \boldsymbol{\Phi}^{-1}\left(F_Y(y)\right)$, R is the correlation coefficient matrix of $\left(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n\right)$ and r is the correlation coefficient vector between $\hat{Y}$ and $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$.

With $\begin{pmatrix} R & r^{\mathrm{T}} \\ r & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \left(R - r^{\mathrm{T}}r\right)^{-1} & -\left(R - r^{\mathrm{T}}r\right)^{-1}r^{\mathrm{T}} \\ -r\left(R - r^{\mathrm{T}}r\right)^{-1} & 1 + r\left(R - r^{\mathrm{T}}r\right)^{-1}r^{\mathrm{T}} \end{pmatrix}$, the log-likelihood function can be rewritten as

$$
\begin{aligned}
\ln L(y) = -\frac{1}{2}\Big\{ & \hat{x}\left[\left(R - r^{\mathrm{T}}r\right)^{-1} - R^{-1}\right]\hat{x}^{\mathrm{T}} \\
& -2r\left(R - r^{\mathrm{T}}r\right)^{-1}\hat{x}^{\mathrm{T}}\hat{y} + r\left(R - r^{\mathrm{T}}r\right)^{-1}r^{\mathrm{T}}\hat{y}^2 \Big\}.
\end{aligned}
$$

MLE y will be derived with equation as below.

$$
\begin{aligned}
\frac{\mathrm{d}\ln L(y)}{\mathrm{d}y} = \frac{f'^{(y)}}{f(y)} + \frac{f(y)}{\varphi\left(\boldsymbol{\Phi}^{-1}(F(y))\right)} \\
r\left(R - r^{\mathrm{T}}r\right)^{-1}\left(\hat{x}^{\mathrm{T}} - r^{\mathrm{T}}\boldsymbol{\Phi}^{-1}(F(y))\right) = 0.
\end{aligned}
$$

If Y follows normal distribution $N(\mu, \sigma)$, the above equation can be simplified as

$$
-\frac{y - \mu}{1} + r\left(R - r^{\mathrm{T}}r\right)^{-1}\left(\sigma\hat{x}^{\mathrm{T}} - r^{\mathrm{T}} \cdot (y - \mu)\right) = 0.
$$

Then

$$
\hat{y} = \mu + \frac{\sigma r\left(R - r^{\mathrm{T}}r\right)^{-1}\hat{x}^{\mathrm{T}}}{1 + r\left(R - r^{\mathrm{T}}r\right)^{-1}r^{\mathrm{T}}},
$$

which is the value $\hat{y}$ as we got for linear regression.

5 Conclusion

In this study, we proposed a new method to estimate CDF and PDF based on a new spline regression, in which the spline function is not always defined by polynomial functions or B-splines, but can be set freely and consists of totally different types of functions in each segment. In this method, the PDF is expressed by piecewise functions instead of series, and with the increase of sample size, the estimated accuracy increases but the complexity of function does not increase. This method is suitable for most types of continuous distributions, and the form of spline function and other parameters does not need to be changed unless the distribution is quite special. The estimation is accurate for various types of distributions and is superior to kernel density estimation. The PDF is always smooth and is not influenced by parameters. The values of estimated CDF are less than 1, positive and monotone increasing. The values of estimated PDF are positive and the integration of PDF is about 1. And it is easy to find a subset from the large sample to reduce the running time and get similar accuracy simultaneously. PDF estimation of high dimensional random variables was also discussed and its potential application in Bayesian classification models and maximum likelihood regression models was presented.

Acknowledgements This work are supported by grants from the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National High-Tech R&D Program of China (863 Program Contract no. 2012AA020307), the National Basic Research Program of China (973 Program) (Contract no. 2012CB721000), and Ph.D. Programs Foundation of Ministry of Education of China (Contract no. 20120073110057), the Young Scholars (Grant no. 31400704) of Natural Science Foundation of China, also computing resources provided by Center for High Performance Computing, Shanghai Jiao Tong University.

References

1. Candy JV (2009) Bayesian signal processing: classical, modern and particle filtering methods. Wiley-Interscience, New York
2. Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, New York
3. Mitchell TM, Carbonell JG, Michalski RS (1986) Machine learning: a guide to current research. Kluwer Academic Publishers, Norwell
4. Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, 3rd edn. McGraw-Hill Education, New York
5. Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27(3):832–837
6. Parzen E (1962) On estimation of probability density function and mode. Ann Math Stat 33(3):1065–1076
7. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
8. Terrell GR (1990) The maximal smoothing principle in density estimation. J Am Stat Assoc 85(410):470–477
9. Alexandre LA (2008) A solve-the-equation approach for unidimensional data kernel bandwidth selection. University of Beira Interior, Covilhã
10. Jeon B, Landgrebe DA (1994) Fast Parzen density estimation using clustering-based branch andbound. IEEE Trans Pattern Anal Mach Intell 16(9):950–954
11. Babich GA, Camps OI (1996) Weighted Parzen windows for pattern classification. IEEE Trans Pattern Anal Mach Intell 18(5):567–570
12. Girolami M, He C (2003) Probability density estimation from optimally condensed data samples. IEEE Trans Pattern Anal Mach Intell 25(10):1253–1264

13. Bowers NL (1966) Expansion of probability density functions as a sum of gamma densities with applications in risk theory. Trans Soc Actuar 18 PT.1(52):125–147
14. Van Khuong H, Kong HY (2006) General expression for pdf of a sum of independent exponential random variables. IEEE Commun Lett 10(3):159–161
15. Schwartz SC (1967) Estimation of probability density by an orthogonal series. Ann Math Stat 38(4):1261–1265
16. Engel J (1990) Density estimation with Haar series. Stat Probab Lett 9(2):111–117
17. Vannucci M (1998) Nonparametric density estimation using wavelets; Discussion Paper 95–26, ISDS. Duke University, Durham
18. Howard RM (2010) PDF estimation via characteristic function and an orthonormal basis set. In: Wseas international conference on systems
19. Xie J, Wang Z (2009) Probability density function estimation based on windowed fourier transform of characteristic function. In: International congress on image and signal processing
20. Wold S (1974) Spline functions in data analysis. Technometrics 16(1):1–11
21. Reinsch CH (1967) Smoothing by spline functions. Numer Math 10(3):177–183
22. Marsh L, Cormier DR (2002) Spline regression models. J R Stat Soc 52(3):49–58
23. Zong Z, Lam KY (1998) Estimation of complicated distributions using B-spline functions. Struct Saf 20(4):341–355
24. Mansour A, Mesleh R, Aggoune EHM (2015) Blind estimation of statistical properties of non-stationary random variables. J Adv Signal Process 51(1):309–314
25. Kitahara D, Yamada I (2015) Probability density function estimation by positive quartic C 2 -spline functions. In: IEEE international conference on acoustics, speech and signal processing
26. De Boor C (1978) A practical guide to splines. Springer, New York
27. Zhang H (2005) The optimality of Naive Bayes. In: Seventeenth international florida artificial intelligence research society conference, Miami Beach, Florida, USA
28. Rennie JDM, Shih L, Teevan J, Karger D (2003) Tackling the poor assumptions of Naive Bayes text classifiers. In: Proceedings of the twentieth international conference on machine learning, Washington, DC, USA
29. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning, Pittsburgh, PA, USA
30. Pfanzagl J, Hamböker R (1996) Parametric statistical theory. J Am Stat Assoc 91(433):269–287